Journal of Agriculture, Biology and Applied Statistics
Vol. 4, Nos. 1-2, 2025, pp. 17-25

URL : www.arfjournals.com
https://DOI:10.47509/JABAS.2025.v04i1-2.03

Forecasting Marigold Production Using Machine Learning Techniques in Chitradurga District of Karnataka

Lalita V Meli[1] and Vasantha Kumari J.[2]

[1]M. Sc. Scholar. Department of Agricultural Statistics, College of Agriculture, University of Agricultural Sciences, Dharwad, India
[2]Assistant Professor, Department of Agricultural Statistics, University of Agricultural Sciences, Dharwad, India
E-mail: lalitameli50272@gmail.com / vasanthakumarij@uasd.in

To cite this article

Lalita V Meli & Vasantha Kumari J. (2025). Forecasting Marigold Production using Machine Learning Techniques in Chitradurga District of Karnataka. Vol. 4, Nos. 1-2, pp. 17-25.

Abstract: Accurate forecasting of flower crop production is essential for data-driven decision-making and sustainable planning in the floriculture sector. Marigold, a commercially valuable crop in the Chitraduga district of Karnataka, is highly influenced by fluctuating agro-climatic conditions, making production forecasting a vital tool for growers and policymakers. This study presents a comparative analysis of two machine learning algorithms Support Vector Regression (SVR) and k-Nearest Neighbors (KNN) for forecasting Marigold production using historical production data. Model performance was assessed using key evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Logarithmic Error (MSLE), and the Coefficient of Determination (R^2). Empirical results demonstrated that the KNN model outperformed SVR, achieving a lower MAE (408.9535), RMSE (714.1817), and MSLE (0.2340), alongside a substantially higher R^2 value (0.4282), indicating robust predictive accuracy. In comparison, the SVR model recorded higher MAE (411.8339), RMSE (775.2162), MSLE (0.2986), and a lower R^2 (0.3262), reflecting a relatively weaker fit. The superior performance of KNN may be attributed to its ability to capture local nonlinear variations more effectively, which is crucial in heterogeneous agricultural environments. This study underscores the importance of algorithm selection in crop production modeling and highlights the potential of KNN as a reliable tool for forecasting Marigold production. These findings offer practical implications for improving precision agriculture practices and call for further validation of machine learning models across varying agro-ecological contexts to support resilient and informed floricultural development.

Keywords: Chitradurga, Marigold, Support Vector Regression, k-Nearest Neighbors, Forecast, Coefficient of Determination.

Introduction

Forecasting crop production is essential for effective planning, market supply management, and policy-making. Traditional forecasting methods often rely on historical trends, but they

Article History Received : 10 January 2025; Revised : 12 February 2025; Accepted : 19 February 2025; Published : 02 June 2025

may not capture the complex, nonlinear patterns of agricultural data influenced by climatic variability, soil conditions, and management practices. Machine learning (ML) techniques, which can learn patterns from large datasets and handle nonlinearity, offer a promising approach to improve the accuracy of production forecasts. Applying machine learning models for Marigold production in Chitradurga can help farmers, stakeholders, and policymakers make informed decisions, optimize resource allocation, and enhance profitability. This study focuses on analyzing historical production data and developing predictive models using machine learning techniques to forecast Marigold production trends in the district.

Floriculture has witnessed significant growth in Chitradurga district of Karnataka, with Marigold emerging as one of the most important commercial flower crops. Marigold holds a prominent place in Karnataka's floriculture industry due to its vibrant color, long shelf life, and extensive use in religious ceremonies, festivals, and ornamental landscaping. Globally, India is recognized as a leading producer of loose flowers, with about 2.85 lakh hectares under floriculture, yielding 22.84 lakh metric tonnes of loose flowers and 9.47 lakh metric tonnes of cut flowers during 2023-24, of which Marigold contributes approximately 2.15 lakh tonnes (H"9.4% of total loose flower production) (NHB, 2024). Karnataka is among the major floriculture-producing states in India, with nearly 29,700 hectares under flower cultivation, producing 3.47 lakh metric tonnes of commercial flowers in 2022-23, where Marigold alone accounts for about 4.97% of the total flower production (Karnataka State Department of Horticulture, 2023).

Chitradurga district is an important floriculture hub, benefiting from well-drained red soils, moderate rainfall, and proximity to major markets. The crop thrives in sandy loam soils under warm temperatures (20-35°C), and local farmers have increasingly adopted modern practices such as drip irrigation, organic inputs, and improved seed varieties. Despite these advances, production is influenced by climate variability, seasonal price fluctuations, and pest infestations, which can impact farmer profitability.

Methodology

Chitradurga district, located in the central part of Karnataka at approximately 14°132 N latitude and 76°252 E longitude, covers an area of 8,440 sq. km. It is bordered by Davanagere, Chikkamagaluru, Tumkur, and Bellary districts. The terrain is characterized by rocky hills, ridges, and fertile plains, with soils ranging from red sandy loam to black clay. Situated at an average elevation of about 700 m above sea level, Chitradurga experiences a tropical semi-arid climate with hot summers (32-40°C), a monsoon season (June-September), and mild winters (12-28°C). Annual rainfall ranges from 550 mm in the drier regions to over 800 mm in some parts, predominantly from the southwest monsoon, with minor contributions from the northeast monsoon.

This study is based on secondary data on Marigold production (in Metric Tons) in Chitradurga district from 1981-82 to 2022-23. The data were collected from the District Statistical Office, Chitradurga, and the Directorate of Economics and Statistics, Bangalore.

Descriptive Analysis

(a) Mean: The arithmetic mean is defined as the sum of all the observations divided by the total number of observations.

$$\text{Arithmetic mean } (\bar{x}) = \frac{\Sigma_{i=1}^{n} x_i}{n},$$

where,

x_i = i^{th} observation and n = Number of observations.

(b) Coefficient of Variation (%)

Coefficient of Variation:

$$CV = \frac{\sigma}{x} \times 100,$$

where,

x = mean and σ = standard deviation.

Standard deviation:

$$\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2},$$

where,

x_i = i^{th} observation,

x = mean,

n = Number of observations.

Machine Learning Tools

K-Nearest Neighbour

Cover & Hart introduced K-Nearest Neighbour, which is a machine learning method used for regression as well as classification. K-NN considers each data record as a vector in an m-dimensional space (where m is the number of features) and predicts the value of each new sample based on the values of K records that are closest to that point in that space (Enas and Choi, 1986).

How the algorithm decides which of the points from the training set are similar enough to be considered when choosing the class to predict for a new observation is to pick the K closest data points to the new observation and to take the most common class among these (Karthikeya *et al.*, 2020). This is why it is called the K-Nearest Neighbors algorithm. This distance is calculated using various measures such as Euclidean distance, Minkowski distance, and Mahalanobis distance. The larger is K; the better is classification. For instance, the

closeness of the new point x and the training point x_i is measured by a Euclidean distance function in the form of the equation as follows

$$d(x, x_i) = \sqrt{\Sigma_{j=1}^{m} \left(x_i^j - x^j \right)^2}, \quad i = 1, 2..., n. \text{ and } j = 1, 2..., m.$$

Where *n* is the number of training samples and *m* is the number of input samples. Samples that are closer to the new sample will have a greater impact on the prediction.

The following are the steps to be followed while working with the K-NN algorithm:

1. **Data pre-processing**: In this step, the dataset is cleaned to remove any blank or unnecessary data. The data is then normalized to guarantee that all features are on the same scale. The most pertinent features for the model can also be chosen using feature selection approaches.
2. **Choosing the value of K**: The value of K is the number of neighbor that will be considered when making a prediction. A small value of K may result in overfitting, while a large value may result in underfitting. The value of K is chosen based on the nature of the problem and the characteristics of the data. This can be done through trial and error or by using cross-validation techniques.
3. **Model training**: The entire dataset is split into training and testing data. Once the value of K is selected, the K-NN model is trained on the training dataset. The goal is to find the K nearest neighbors for each data point in the dataset. This is done by calculating the Euclidean distance between each data point and all the other data points in the dataset. The K nearest neighbors are the data points that are closest to the input data point.
4. **Model validation**: The model is then tested on a different test dataset after training. Metrics including RMSE, MAE, MSLE and R^2 are used to assess the model's performance. This is carried out to make sure the model is not overfitting and to determine how well it will function with new data.
5. **Model optimization**: If the model is not performing well, the value of K can be adjusted to improve performance. Grid search and cross-validation techniques can be used to find the best value of K.
6. **Model deployment**: Once the model has been improved, it can be applied to predict new data. The trained model receives the input attributes, predicts the output value based on the majority class, and outputs that value.

3.3.2.2. Support Vector Regression

Support Vector Regression was introduced by Vladimir *N*. Vapnik and Alexey Ya. Chervonenkis in 1963, is the regression model of Support Vector Machine, on a dataset consisting of *L* samples of form $\{(x_1, y_2), (x_1, y_2), \ldots, (x_L, y_L), x \in R^m, y \in R\}$ is a linear function which can estimate output values based on inputs.

$$y = (w. x) + b,$$

where y is the estimated value, x is the input vector, w is the weight vector and b is the bias.

SVR creates a hyperplane or sets of hyperplanes in a high or infinite-dimensional space, which is utilized for regression, classification, or other tasks. SVR uses linear functions for learning. In the case of nonlinear cases, SVR uses a kernel technique to plot the data into a higher-dimensional feature space, in which linear functions can be applied (Palanivel and Suriyanarayanan, 2019).

Hyperparameter in SVR

1. **Hyperplane**: Hyperplanes are decision boundaries for predicting the continuous output. Support Vectors are the data points on either side of the hyperplane that are closest to the hyperplane. These are used to draw the required line that shows the algorithm's predicted outcome.
2. **Kernel**: A kernel is a collection of mathematical functions that take data and change it into the desired form. These are most commonly used to find a hyperplane in higher-dimensional space. Linear, Non-Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid are the most commonly used kernels. RBF is the kernel that is used by default. Depending on the dataset, each of these kernels is used.
3. **Boundary Lines**: These are the two lines that are drawn at a distance of e (epsilon) from the hyperplane. It's used to separate the data points by a margin.
4. **Support Vectors**: The closest point of the lines from both classes.

   The following are the steps to be followed while working with the SVR algorithm:

   1. **Data pre-processing**: In this step, the dataset is cleaned to remove any missing or irrelevant data. The data is then normalized to ensure that all features are on the same scale. Feature selection techniques can also be used to select the most relevant features for the model.
   2. **Kernel selection**: A kernel function is a mathematical function that transforms the data into a higher-dimensional space. The kernel function is chosen based on the nature of the problem and the characteristics of the data. Some popular kernels are linear, polynomial, and radial basis functions.
   3. **Model training**: The entire dataset is split into training and testing data and the SVR model is trained on the training dataset. The goal is to find the optimal hyperparameters that maximize the margin between the predicted and actual values. The margin is the distance between the hyperplane and the closest data points. The hyperparameters that are tuned include the regularization parameter (C), the kernel coefficient (gamma), and the kernel type.

4. **Model validation**: After training the model, it is validated on a separate test dataset. The performance of the model is evaluated using metrics such as RMSE, MAE, MSLE and R^2. This is done to ensure that the model is not overfitting and to get an idea of how well it will perform on new data.
5. **Model optimization**: If the model is not performing well, the hyperparameters can be adjusted to improve performance. Grid search and cross-validation techniques can be used to find the best values for the hyperparameters, especially in the case of smaller datasets.
6. **Model deployment**: Once the model is optimized, it can be used to make predictions on new data. The trained model takes in the input features and returns the predicted output value.

Results and Discussion

The mean Marigold production in Chitradurga was 1380.14 MT, with a high SD of 955.61 and CV of 69.24%, indicating considerable year-to-year variability. The positive skewness (1.37) shows occasional years of exceptionally high yields, while the kurtosis (2.48) suggests a moderately flat distribution with fewer extreme values. These statistics highlight the unpredictability of Marigold production, underscoring the need for forecasting models to support planning and risk management, as shown in Table 1. These results are on par with Dwivedi *et al.* (2024) and Evangilin *et al.* (2020).

In developing a machine learning architecture for forecasting the production of major flower crops in the selected districts, two distinct models were implemented: Support Vector Regression SVR and K-nearest neighbour (K-NN). The algorithm was written in the Visual Studio (VS) Code environment using Python. To enhance the model's accuracy, feature selection and cross-validation techniques were employed.

The effectiveness of both models was assessed using the following statistical evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Logarithmic Error (MSLE), and Coefficient of determination (R^2).

As shown in Table 2, the performance of two machine learning algorithms Support Vector Regression (SVR) and K-Nearest Neighbors (K-NN) was evaluated for forecasting Marigold production in Chitradurga district using MAE, RMSE, MSLE, and R^2 as performance metrics. K-NN outperformed SVR across all measures, achieving a lower MAE of 408.95 compared to 411.83 for SVR, indicating that its predictions were, on average, closer to the actual production values. Similarly, K-NN recorded a lower RMSE of 714.18 against SVR's 775.22, reflecting its better capability to handle years with large fluctuations in production. The MSLE for K-NN was 0.23, lower than SVR's 0.30, demonstrating that K-NN more accurately captured proportional differences in production, which is crucial given the variability in agricultural yields. Furthermore, the coefficient of determination (R^2) was higher for K-NN (0.43) than for SVR (0.33), showing that K-NN explained a

larger proportion of the variability in Marigold production. Overall, these results indicate that K-NN provides more accurate and reliable forecasts, closely following actual production trends, especially during years with sharp increases or decreases, whereas SVR struggled to capture such fluctuations. The superior performance of K-NN can be attributed to its ability to leverage local patterns in historical data, making it particularly effective for forecasting agricultural production with inherent variability.

As shown in Table 3, by using best fitted model, forecasted the production values for Marigold in Chitraduraga for the next five years 2023-28. Figure 1 shows the actual and forecasted production values by using best best-fitted model. actual Marigold production initially showed a consistent rise until 2000, supported by favourable rainfall and expansion of cultivation under open conditions. However, 2001 marked the beginning of a declining trend in production, attributed to delayed monsoon onset and insufficient flowering due to heat stress during the vegetative phase, as reported in IMD Karnataka district-level seasonal summaries (Anon., 2001, IMD). The massive spike in 2011 was linked to district-level convergence programs under RKVY, leading to a sharp increase in area and output. Again in 2015 due to high market value the production increased in that year also. The results are

Table 1: Descriptive Statistics of Marigold Production in Chitradurga District of Karnataka

Parameter	*Descriptives*	*Marigold*
Production	Mean	1380.14
	SD	955.61
	CV (%)	69.24
	Skewness	1.37
	Kurtosis	2.48

Table 2: Comparison of machine learning algorithms for the forecast of Marigold production in Chitradurga

Algorithm	*MAE*	*RMSE*	*MSLE*	R^2
SVR	411.83	775.22	0.30	0.33
K-NN	408.95	714.18	0.23	0.43

Table 3: Forecasted Production Values for Marigold in Chitradurga District for the Next Five Years Based on the Best-Fitted Model

Year	*Production (MT) using K-NN model*
2024	2221.12
2025	2458.12
2026	2695.14
2027	2932.17
2028	3169.21

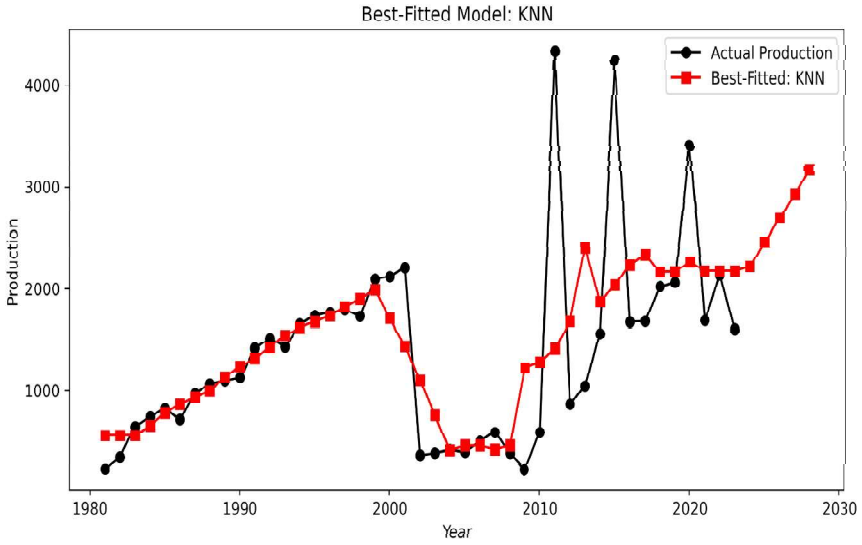

Fig. 1: Forecast of production of the Marigold flower crop in Chitradurga district using the best-fitted model

consistent with earlier findings by Akhila *et al.* (2023) and Bondre & Mahagaonkar (2019), who demonstrated the superior performance of K-NN in modeling non-linear agricultural data.

References

1. Akhila P S, Ashalatha K V, Vasantha Kumari J, Milind P P and Satish R D,2024, A data-driven approach to predict crop yield using AI tools. *M.Sc. (Agri) Thesis*, University of Agricultural Sciences, Dharwad, Karnataka, India.
2. Anonymous, 2001, Ministry of Earth Sciences, Government of India. Annual Climate Summary - Karnataka Region. mausam.imd.gov.in.
3. Bondre D A and Mahagaonkar S, 2019, Prediction of crop yield and fertilizer recommendation using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology*, 4(5): 371-376.
4. Dwivedi R K, Khavse R and Ahriwar M K, 2024, Trend analysis of area, production and productivity of small millets in Madhya Pradesh, India. *Plant Archives*, 24(1): 543-548.
5. Enas G G and Choi S C, 1986, Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. *Statistical Methods of Discrimination and Classification*, 2(3): 235-244.
6. Evangilin N P, Murthy B R, Naidu M, and Aparna B, 2020, Statistical model for forecasting area, production and productivity of Sesame crop (Sesamum indicum L.) in Andhra Pradesh, India. *International Journal of Current Microbiology and Applied Sciences*, 9(7): 1156-1166.

7. Karthikeya H K, Sudarshan K and Shetty D S, 2020, Prediction of agricultural crops using KNN algorithm. *International Journal of Innovative Science and Research Technology*, 5(5): 1422-1424.

8. Palanivel K and Suriyanarayanan C, 2019, An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology*, 10(3): 110-118.